Power Law for Estimating Underdetection of Foodborne Disease Outbreaks, United States

Laura Ford,[1] Julie L. Self,[1] Karen K. Wong, Robert M. Hoekstra, Robert V. Tauxe, Erica Billig Rose, Beau B. Bruce

We fit a power law distribution to US foodborne disease outbreaks to assess underdetection and underreporting. We predicted that 788 fewer than expected small outbreaks were identified annually during 1998–2017 and 365 fewer during 2018–2019, after whole-genome sequencing was implemented. Power law can help assess effectiveness of public health interventions.

Each year in the United States, >800 foodborne outbreaks are reported, causing >14,000 illnesses and >800 hospitalizations (*1*–*3*). Foodborne outbreaks range from small, localized outbreaks, such as those associated with a locally contaminated meal shared by family or friends, to large, multistate outbreaks associated with a contaminated food that is widely distributed. Selection and information biases, pathogen testing methods, and outbreak size can affect detection, investigation, and reporting (*4*). However, few methods are available to estimate the extent of outbreak underdetection and underreporting.

Outbreaks can be considered natural occurrences with a mathematical relationship between frequency and size. Several studies have used a power law distribution, where one variable is proportional to the power of another, to help describe disease outbreaks or transmission (*5*–*9*). We examined the mathematical relationship between foodborne outbreak frequency and size to estimate the number of expected outbreaks of different sizes, comparing power law, log-normal, and exponential distributions by using censored and complete data to clarify underdetection and underreporting.

The Study

Local, state, and federal public health agencies in the United States identify and investigate foodborne outbreaks and report them to the Foodborne Disease Outbreak Surveillance System (FDOSS; https://www.cdc.gov/fdoss). In FDOSS, a foodborne outbreak is defined as ≥2 similar illnesses associated with a common food source. We used FDOSS data from 1998–2019 and defined outbreak size as the number of laboratory-confirmed cases. We also included outbreaks with ≥2 similar illnesses that had only 1 confirmed case. We evaluated the fit of power law, log-normal, and exponential distributions by applying the Kolmogorov-Smirnov (KS) statistic (*10*) to the number of outbreaks by size.

We estimated medians and 90% credible intervals (CrIs) for the minimum threshold, slope, and difference between expected and actual outbreak frequency by bootstrapping 5,000 random samples with replacement from the dataset of all outbreaks of the same size. We defined outbreaks of <10 confirmed cases as small and outbreaks of >100 confirmed cases as large. We conducted all analyses in R (The R Foundation for Statistical Computing, https://www.r-project.org) by using the poweRlaw package version 0.70.6 (*11*). We provide additional methods and R script (Appendix 1, https://wwwnc.cdc.gov/EID/article/30/2/23-0342-App1.pdf) and the dataset used (Appendix 2, https://wwwnc.cdc.gov/EID/article/30/2/23-0342-App2.xlsx).

During 1998–2019, a total of 10,026 foodborne outbreaks were reported in the United States, ranging from 1 to 1,500 laboratory-confirmed cases. The data appeared linear on a log-log scale, consistent with a power law distribution (Figure 1, panel A). We rejected the exponential and log-normal distributions because they fit poorly based on the KS statistic (exponential 0.109, p<0.001; log-normal 0.0101, p<0.001). The power law distribution fit the data (KS = 0.00985, p = 0.15).

Foodborne outbreaks with ≥4 (90% CrI 4–8) cases followed a power law distribution of $\alpha = 2.15$ (90%

Author affiliation: Centers for Disease Control and Prevention, Atlanta, Georgia, USA

DOI: https://doi.org/10.3201/eid3002.230342

[1]These first authors contributed equally to this article.

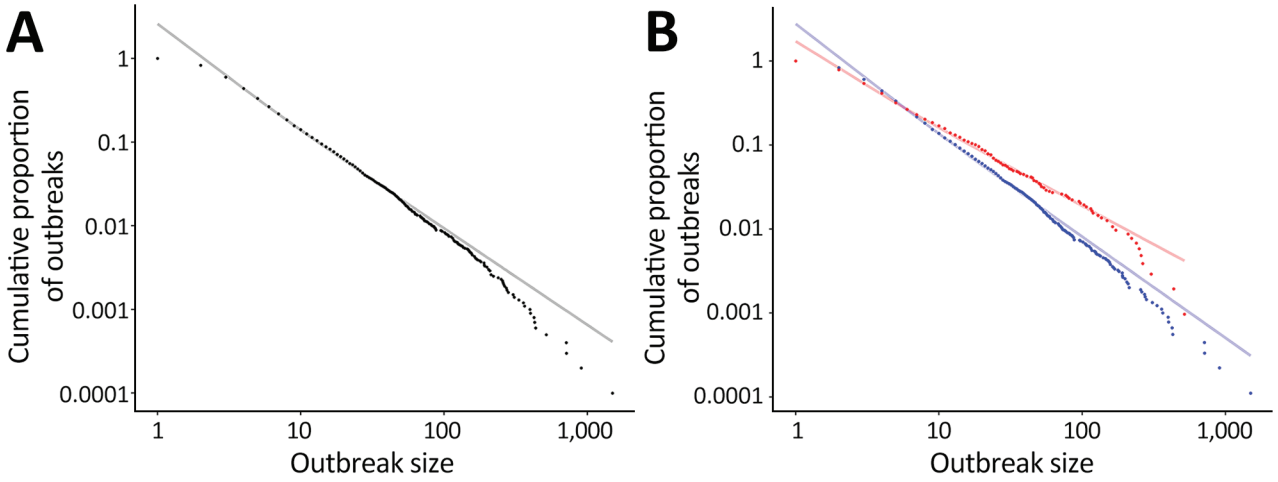

Figure 1. Log-log scale of foodborne outbreak size versus frequency from a power law for estimating underdetection of foodborne disease outbreaks, United States. A) Actual (black points) versus expected from the power law distribution (gray line) 1998–2019; B) actual (blue points) versus expected (light blue line) 1998–2017 and actual (red points) versus expected (light red line) 2018–2019. Estimates for the difference between the number of expected and actual small (<10 cases) and large (>100 cases) outbreaks were calculated by the sum of the differences between each of the relevant actual points and the expected line at the same x-value. Annual estimates were then calculated by dividing the number of years represented.

CrI 2.12–2.19) (Figure 2). We estimated 718 (90% CrI 594–783) fewer than expected small outbreaks and 0.4 (90% CrI −0.07–0.9) fewer than expected large outbreaks occurred annually, representing 841 (90% CrI 669–932) fewer than expected small outbreak-associated illnesses and 574 (90% CrI 325–871) fewer than expected large outbreak-associated illnesses.

By 2018, most US public health laboratories were using whole-genome sequencing (WGS) to subtype some bacteria that cause foodborne illness, including *Salmonella enterica*, *Escherichia coli*, and *Listeria monocytogenes*. WGS has helped public health practitioners detect more outbreaks and determine the food or other source while outbreaks are still small (*12*).

A power law distribution fit the outbreak data for both the 1998–2017 (8,993 outbreaks; KS = 0.00949, p = 0.37) and the 2018–2019 (1,033 outbreaks; KS = 0.0211, p = 0.43) periods (Figure 1, panel B). The minimum threshold was ≥5 cases (90% CrI 4–9) and α = 2.20 (90% CrI 2.16–2.25) during 1998–2017, compared with

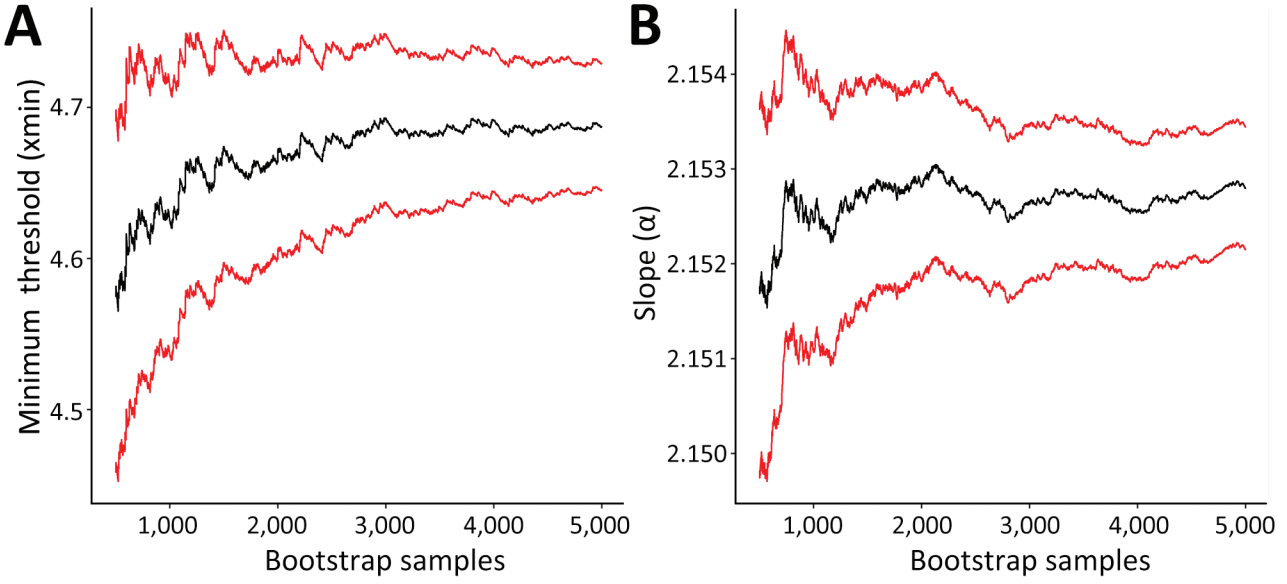

Figure 2. Parameter estimates from a power law for estimating underdetection of foodborne disease outbreaks, United States. Graphs display distribution of foodborne outbreak size and frequency for the minimum threshold (A) and slope (B) for outbreaks during 1998–2019. Black lines represent bootstrapped parameter estimate; red lines represent 90% credible intervals.

a minimum threshold of ≥3 cases (90% CrI 2–6) and α = 1.91 (90% CrI 1.83–2.00) during 2018–2019. We estimate 788 (90% CrI 665–888) fewer than expected small outbreaks and 0.4 (90% CrI −0.06 to 0.9) fewer than expected large outbreaks were identified annually during 1998–2017, compared with 365 (90% CrI 277–475) fewer than expected small outbreaks and 1 (90% CrI −3 to 2) more than expected large outbreak annually during 2018–2019.

Conclusions

We found that foodborne disease outbreak data fit a power law distribution. On the basis of that finding, we quantified the unobserved burden of foodborne outbreaks in the United States during 1998–2019, predicting that 718 fewer than expected small outbreaks are detected, investigated, and reported every year and 1 fewer than expected large outbreak was detected and reported about every 3 years. Detection and reporting of foodborne outbreaks have improved; during 2018–2019, we estimate that underreporting of small outbreaks decreased by 54% (365/year) compared with 1998–2017 (788/year). The power law distribution quantifies improvements in detection and reporting, which could in part be explained by WGS.

Many factors affect outbreak and case detection, investigation, and reporting, including whether the outbreak is caused by a common molecular strain, how many persons ate the contaminated food, clinical manifestations, care-seeking, diagnostic testing, and laboratory or health department outbreak investigation and response capacity. Natural limitations to outbreak size are also likely, including the geographic distribution of a contaminated food product, food safety policies that control contamination in the food system, and product recalls or other disease control efforts that end large outbreaks before natural limitations are reached.

Power law distribution parameters should be stable over time, but changes in the slope or minimum threshold or deviations from the estimated power law might indicate perturbations of concern. Understanding the different power law parameters that underlie outbreak size and frequency can also be useful for exploring how detection of foodborne outbreaks differs by pathogen or food vehicle. In addition, those parameter changes can reflect public health interventions.

The power law distribution has applications beyond foodborne outbreaks and has been applied to COVID-19, measles, and gonorrhea (*5*–*9*). By predicting outbreak frequency and the extent of underdetection, we can plan outbreak response needs for routine and surge scenarios, assess the effects of outbreak prevention efforts, and improve estimates of the proportion of illnesses that are outbreak-associated versus sporadic.

A limitation of this analysis is that failure to statistically reject the power law distribution does not ensure that the data follow a power law. The KS statistic also might miss systematic patterns that differ between distributions because it uses only the largest difference. However, we used a hypothesis-driven rationale to censor data by establishing a minimum threshold, tested alternative distributions, and characterized uncertainty by using the bootstrap. Another limitation is that we only include reported outbreaks with laboratory confirmed cases, which could underestimate cases but also reduces variation from comparing across multiple types of outbreaks. Laboratory-confirmed cases also could be an underestimate for the largest outbreaks because public health laboratories might run out of resources to subtype patient samples or be faced with other constraints due to the overwhelming size of the outbreak.

In conclusion, we used the power law distribution on foodborne disease outbreak data to quantify underdetection and how foodborne disease reporting has improved. The improvement in underdetection during 2018–2019 could in part be explained by improved detection or investigation from the implementation of WGS. The power law distribution can be used to assess the impact of past and future public health interventions and as a tool for resource planning.

This work was supported by the Centers for Disease Control and Prevention. The study did not receive dedicated funding.

About the Author

Dr. Ford is an epidemiologist in the Division of Foodborne, Waterborne, and Environmental Diseases, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA. Her primary research interests include surveillance and outbreak response for foodborne diseases.

References

1. Centers for Disease Control and Prevention (CDC). Surveillance for foodborne disease outbreaks, United States, 2015, annual report. Atlanta: US Department of Health and Human Services, CDC; 2017.
2. Centers for Disease Control and Prevention (CDC). Surveillance for foodborne disease outbreaks, United States, 2016, annual report. Atlanta: US Department of Health and Human Services, CDC; 2018.
3. Centers for Disease Control and Prevention (CDC). Surveillance for foodborne disease outbreaks, United States,

2017, annual report. Atlanta: US Department of Health and Human Services, CDC; 2019.
4. Mouly D, Goria S, Mounié M, Beaudeau P, Galey C, Gallay A, et al. Waterborne disease outbreak detection: a simulation-based study. Int J Environ Res Public Health. 2018;15:1505. https://doi.org/10.3390/ijerph15071505
5. Beare BK, Toda AA. On the emergence of a power law in the distribution of COVID-19 cases. Physica D. 2020;412:132649. https://doi.org/10.1016/j.physd.2020.132649
6. Komarova NL, Schang LM, Wodarz D. Patterns of the COVID-19 pandemic spread around the world: exponential versus power laws. J R Soc Interface. 2020;17:20200518. https://doi.org/10.1098/rsif.2020.0518
7. Blasius B. Power-law distribution in the number of confirmed COVID-19 cases. Chaos. 2020;30:093123. https://doi.org/10.1063/5.0013031
8. Keeling M, Grenfell B. Stochastic dynamics and a power law for measles variability. Philos Trans R Soc Lond B Biol Sci. 1999;354:769–76. https://doi.org/10.1098/rstb.1999.0429
9. Whittles LK, White PJ, Didelot X. A dynamic power-law sexual network model of gonorrhoea outbreaks. PLOS Comput Biol. 2019;15:e1006748. https://doi.org/10.1371/journal.pcbi.1006748
10. Chakravarti IM, Laha RG, Roy J. Handbook of methods of applied statistics, volume I. New York: John Wiley & Sons; 1967.
11. Gillespie CS. Fitting heavy tailed distributions: the poweRlaw package. J Stat Softw. 2015;64:1–16. https://doi.org/10.18637/jss.v064.i02
12. Besser JM, Carleton HA, Trees E, Stroika SG, Hise K, Wise M, et al. Interpretation of whole-genome sequencing for enteric disease surveillance and outbreak investigation. Foodborne Pathog Dis. 2019;16:504–12. https://doi.org/10.1089/fpd.2019.2650

Address for correspondence: Laura Ford, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, MS H24-11, Atlanta, GA 30329-4018, USA; email: qdz4@cdc.gov

Article DOI: https://doi.org/10.3201/eid3002.230342

Power Law for Estimating Underdetection of Foodborne Disease Outbreaks, United States

Appendix 1

Methods

The R (The R Foundation for Statistical Computing, https://www.r-project.org) script used in this analysis is below. The script evaluates the fit of an exponential, log-normal and power law distribution and estimates the median and 90% credible interval for minimum threshold, slope, and difference between expected and actual frequency based on the power law distribution. The estimates for the difference between the number of expected and actual small (<10 cases) and large (>100 cases) outbreaks were calculated by the sum of the differences between each of the relevant actual points and the expected line at the same x-value. Annual estimates are then calculated by dividing the number of years represented.

#The following is the R script used in: Ford L, Self JL, Wong KK, Hoekstra RM, Tauxe RV, Rose EB, Bruce BB. Power Law for Estimating Underdetection of Foodborne Disease Outbreaks, United States . Last modified in 2023.

```
library(tidyverse)
library(poweRlaw)
library(readxl)
library(colorspace)
library(ggpubr)
##############################################################################
##
##
```

```
## Prepare data
##
#############################################################################
##
read_excel("FORD_20220531_forrelease.xlsx,"
sheet = "FORD_20220531_forreview2") %>%
filter(ConfirmedPrimary ! = 0 and yearfirstill ! = 2020) %>%
.$ConfirmedPrimary %>%
na.omit %>%
as.vector ->
d
#############################################################################
##
##
## Explore distribution fits
##
#############################################################################
##
## Discrete exponential distribution
dexp <- disexp$new(d)
dexp$setPars(estimate_pars(dexp))
dexp_p < - bootstrap_p(dexp, xmins = 1:100, no_of_sims = 5000, threads = 4)
plot(dexp_p)
dexp_p$p
## Log normal distribution
dln <- dislnorm$new(d)
dln$setPars(estimate_pars(dln))
dln_p < - bootstrap_p(dln, xmins = 1:100, no_of_sims = 5000, threads = 4)
plot(dln_p)
dln_p$p
## Power law distribution with cutoff
dplco <- displ$new(d)
dplco$setXmin(estimate_xmin(dplco))
```

```
dplco$setPars(estimate_pars(dplco))
dplco_p < - bootstrap_p(dplco, xmins = 1:20, no_of_sims = 5000, threads = 4)
plot(dplco_p)
dplco_p$p
quantile(dplco_p$bootstraps$xmin[1:5000], probs = c(0.05, 0.5, 0.95))
quantile(dplco_p$bootstraps$pars[1:5000], probs = c(0.05, 0.5, 0.95))
##############################################################################
##
##
## Estimate the difference between actual and expected from
## power law distribution overall (1998–2019)
##
##############################################################################
##
# properly aligned cumsum for our purposes
fixcumsum <- function(x) {
x <- rev(x)
for (i in head(seq_along(x), −1)) {
if (is.na(x[i + 1])) {
x[i + 1] <- x[i]
}
}
rev(x)
}
# apply the bootstrap parameters to the distribution object and return expected versus actual
uo2 <- function(dis, α, altxmin) {
oldalpha <- dis$getPars()
dis$setPars(α)
oldxmin <- dis$getXmin()
dis$setXmin(altxmin)
xmin <- altxmin
vals <- dis$internal$values
lvals <- log10(vals)
```

```
cumn <- dis$internal$cum_n
lcumn <- log10(cumn)
plot(lvals, lcumn, pch = 20)
yhat <- dist_cdf(dis, tail(vals, -(xmin - 1)), FALSE)
dis$setXmin(1)
scale <- dist_data_cdf(dis, FALSE)[xmin]
cumn <- dis$internal$cum_n[1] * yhat * scale
lcumn <- log10(cumn)
dis$setXmin(oldxmin)
slope <- diff(head(lcumn, 2)) / diff(tail(lvals, -(xmin - 1)))[1]
icept <- lcumn[1] - slope * log10(xmin)
lcumn <- c(slope * lvals[1:(xmin - 1)] + icept, lcumn)
cumn <- c(10 ^ lcumn[1:(xmin - 1)], cumn)
lines(lvals, lcumn, col = “red”)
# extrapolate to y = 0 using CDF
slope2 <- diff(tail(lcumn, 2)) / diff(tail(lvals))
icept2 <- tail(lcumn, 1) - slope2 * max(lvals)
maxx <- c(lvals, -icept2 / slope2)
xs <- min(vals):max(10 ^ maxx)
lxs <- log10(xs)
lyhats <- approx(c(lvals, max(lxs)), c(lcumn, 0), lxs)$y
yhats <- 10 ^ lyhats
out <- data.frame(xs, lxs, lyhats, yhats)
out <- merge(out, data.frame(xs = dis$internal$values,
yobs = dis$internal$cum_n),
all.x = TRUE)
out$yobs <- fixcumsum(out$yobs)
out$yobs[is.na(out$yobs)] <- 0
out$ydiff <- c(-diff(out$yhats) - -diff(out$yobs),
tail(out$yhats, 1) - tail(out$yobs, 1))
out$ndiff <- out$xs * out$ydiff
dis$setPars(oldalpha)
out
```

```
}
## function to calculate difference in actual versus expected for small and large
## outbreaks
small_big <- function(dis, dis_p) {
out <- list()
boot_slope <- dis_p$bootstraps$pars
boot_xmin <- dis_p$bootstraps$xmin
for (i in 1:5000) {
o <- uo2(dis, boot_slope[i], boot_xmin[i])
# includes confirmed ill of size = 1 to 9
smallobs <- with(o, sum(ydiff[xs <10]))
smallills <- with(o, sum(ndiff[xs <10]))
# includes confirmed ill of size 100 or larger
bigobs <- with(head(o, −1), sum(ydiff[xs >100]))
bigills <- with(head(o, −1), sum(ndiff[xs >100]))
out$diffs$small_obs <- c(out$diffs$small_obs, smallobs)
out$diffs$small_ills <- c(out$diffs$small_ills, smallills)
out$diffs$big_obs <- c(out$diffs$big_obs, bigobs)
out$diffs$big_ills <- c(out$diffs$big_ills, bigills)
}
out$summary <- lapply(out$diffs, quantile, probs = c(0.05, 0.5, 0.95))
out
}
o_small_big <- small_big(dplco, dplco_p)
##############################################################################
##
##
## Fit separate power law distributions to 1998–2017 and 2018–2019 and compare
##
##############################################################################
##
### 1998–2017
read_excel("FORD_20220531_forrelease.xlsx,"
```

```
sheet = “FORD_20220531_forreview2”) %>%
filter(ConfirmedPrimary ! = 0 and yearfirstill <2018) %>%
.$ConfirmedPrimary %>%
na.omit %>%
as.vector ->
bwgs
bwgsplco <- displ$new(bwgs)
bwgsplco$setXmin(estimate_xmin(bwgsplco))
bwgsplco$setPars(estimate_pars(bwgsplco))
bwgsplco_p < - bootstrap_p(bwgsplco, xmins = 1:20, no_of_sims = 5000, threads = 4)
plot(bwgsplco_p)
bwgsplco_p$p
quantile(bwgsplco_p$bootstraps$xmin[1:5000], probs = c(0.05, 0.5, 0.95))
quantile(bwgsplco_p$bootstraps$pars[1:5000], probs = c(0.05, 0.5, 0.95))
boot_slope_bwgs <- bwgsplco_p$bootstraps$pars
boot_xmin_bwgs <- bwgsplco_p$bootstraps$xmin
b_small_big <- small_big(bwgsplco, bwgsplco_p)
## 2018–2019
read_excel(“FORD_20220531_forrelease.xlsx,”
sheet = “FORD_20220531_forreview2”) %>%
filter(ConfirmedPrimary ! = 0 &
(yearfirstill > = 2018 and yearfirstill <2020)) %>%
.$ConfirmedPrimary %>%
na.omit %>%
as.vector ->
awgs
awgsplco <- displ$new(awgs)
awgsplco$setXmin(estimate_xmin(awgsplco))
awgsplco$setPars(estimate_pars(awgsplco))
awgsplco_p < - bootstrap_p(awgsplco, xmins = 1:20, no_of_sims = 5000, threads = 4)
plot(awgsplco_p)
awgsplco_p$p
awgsplco_p$bootstraps
```

```
quantile(awgsplco_p$bootstraps$xmin[1:5000], probs = c(0.05, 0.5, 0.95))
quantile(awgsplco_p$bootstraps$pars[1:5000], probs = c(0.05, 0.5, 0.95))
boot_slope_awgs <- awgsplco_p$bootstraps$pars
boot_xmin_awgs <- awgsplco_p$bootstraps$xmin
a_small_big <- small_big(awgsplco, awgsplco_p)
##############################################################################
##
##
## Figures
##
##############################################################################
##
## Figure 1
options(scipen = 100000)
# function to generate consistent plot data
plot_data <- function(dis) {
xmin <- dis$getXmin()
uo2(dis, dis$getPars(), xmin) %>%
filter(xs %in% dis$internal$values) %>%
mutate(yobs = yobs / yobs[1], yhats = yhats / yhats[xmin] * yobs[xmin])
}
# function to generate a consistent power law plot
ggpl <- function(data) {
if (is.null(data$grp)) data$grp <- "A"
ggplot(data, aes(x = xs, y = yhats, group = grp, color = grp)) +
geom_line(linewidth = 1, α = 0.25) +
geom_point(aes(y = yobs), size = 0.5) +
scale_x_log10(name = "Outbreak size") +
scale_y_log10(name = "Cumulative proportion of outbreaks,"
breaks = 10 ^ (−4:0),
labels = 10 ^ (−4:0)) +
theme_bw()+
theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
```

```
}
ggpl(plot_data(dplco)) +
ggtitle("A.") +
scale_color_manual(values = "black") ->
all_years
plot_data(bwgsplco) %>%
mutate(grp = "1998-2017") %>%
bind_rows(plot_data(awgsplco) %>%
mutate(grp = "2018-2019")) %>%
ggpl() +
ggtitle("B.") +
scale_color_manual(values = c("blue," "red")) ->
comb
tiff(filename = "Figure 1.tiff," height = 5, width = 12,
units = "in," res = 600)
ggarrange(all_years, comb, ncol = 2, legend = "none")
dev.off()
## Figure 2
plot_data2 <- function(dpl) {
dplco_p$bootstraps %>%
select(xmin, pars) %>%
map(poweRlaw:::get_cum_summary) %>%
imap(~ mutate(.x, var = .y)) %>%
map(~ select(., x, var, m, m_up, m_low) %>%
pivot_longer(-c(var, x))) %>%
bind_rows %>%
mutate(var = ifelse(var =  = "pars," "alpha," "X_min"),
var = recode(var, "alpha" = "Slope (\u03b1),"
"X_min" = "Minimum Threshold (xmin)"))
}
tiff(filename = "Figure 2.tiff," height = 5, width = 10, units = "in," res = 600)
plot_data2(dplco_p) %>%
ggplot(aes(x = x, y = value, color = name)) +
```

```
geom_line() +
facet_wrap(~ var, scales = “free”) +
scale_color_manual(values = c(“black,” “red,” “red”)) +
guides(color = “none”) +
xlab(“Bootstrap samples”) +
theme_bw() +
theme(axis.title.y = element_blank(), panel.grid.major = element_blank(), panel.grid.minor =
element_blank())
dev.off()
```